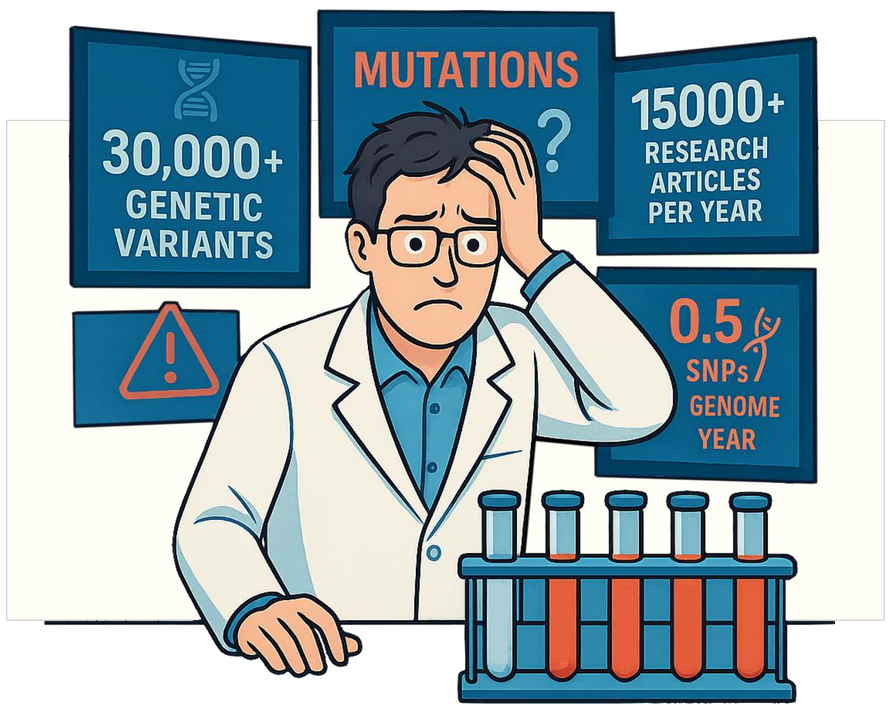
RAG for *Mycobacterium tuberculosis* Data Access

Jihad Al Akl[1,2] , David Laiymani[1], Christophe Guyeux[1], Christophe Sola[3,4], Chady Abou Jaoude[2], and Zahi Al Chami[2]
[1]FEMTO-ST Institute, CNRS, University of Marie & Louis Pasteur,
[2]Ticket Lab, Antonine University, Lebanon
[3]IAME, UMR1137, INSERM, Universite Paris-Cité, Université Sorbonne Paris-Nord
[4]Université Paris-Saclay



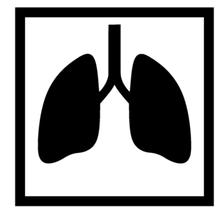
Context & Problems

Every day, more than 3,000 new biomedical research papers are added to websites like PubMed. For scientists studying Mycobacterium—a group of bacteria that includes the pathogen responsible for tuberculosis—the volume of information to keep up with is overwhelming and constantly growing. Thousands of Mycobacterium genomes have already been sequenced, revealing millions of genetic mutations. New studies continue to add insights into gene variations, drug resistance, potential treatments, and vaccine development. The pace of discovery makes it increasingly difficult to stay informed. Researchers often find themselves spending more time keeping up with the literature than conducting their own investigations. Compounding the challenge is the ever-evolving nature of science itself. New research can change what we thought was true or show that things we once thought didn't matter are actually important. For instance, a mutation once deemed irrelevant might later prove critical to understanding how the bacteria evade antibiotics.

Objectives

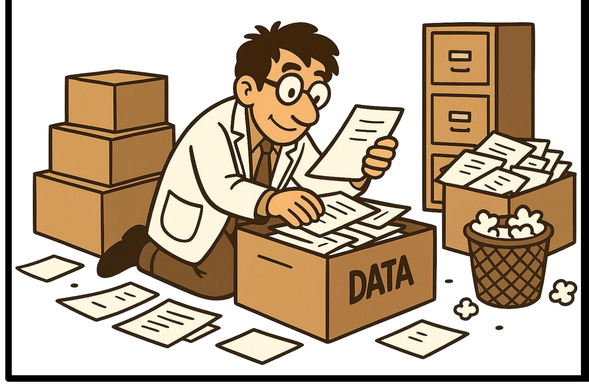
- Develop an AI-powered tool to assist scientists in tracking news within their fields.
- Ensure timely updates with the latest scientific news and developments.
- Design an intuitive, user-friendly platform for seamless experience.
- Continuously update AI models to maintain accuracy and relevance.

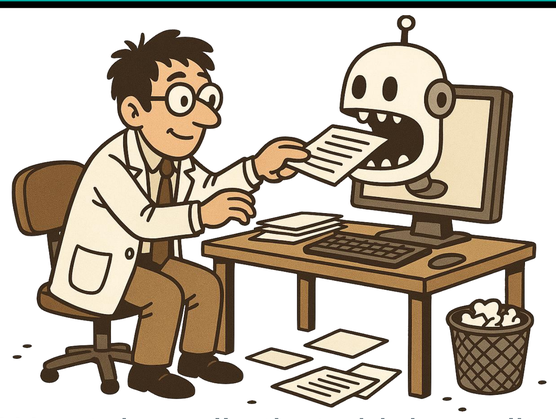
Interact



Methodology

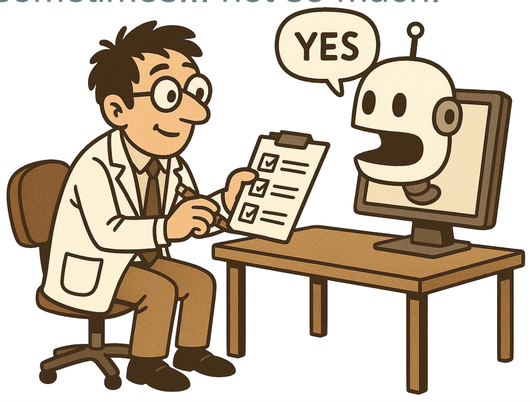
We dig through tons of research, keep what's valuable, and toss the junk into the bin. Only the most reliable data make the cut.

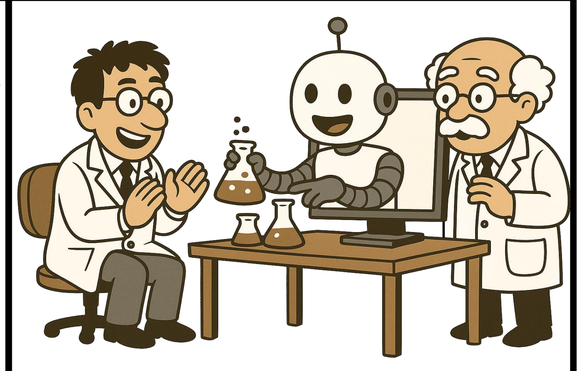




We take all that high-quality data and feed it into our model. It devours information like it's at an all-you-can-read buffet

Time to put our model to the test! We throw questions at it and cross our fingers.
Sometimes it nails the answer, sometimes... not so much.

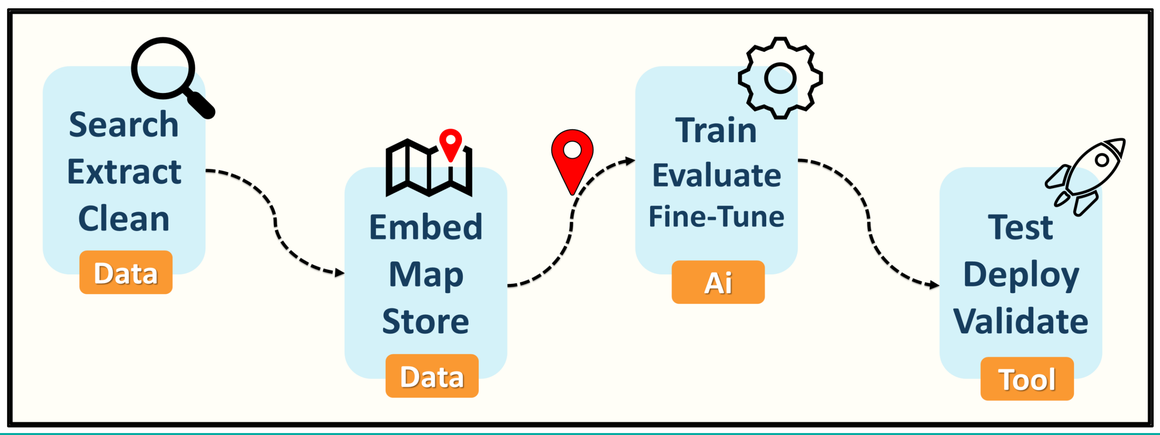




Our model now produces valuable outputs, ideas, and insights. We collaborate with experts for testing, validation, and proper acknowledgment.

Preliminary Results

- Successfully extracting daily medical reports from trusted news sources.
- Identifying and extracting newly reported mutations and associated drugs.
- Generating concise summaries of articles using LLMs techniques.
- Indexing and organizing extracted information in a searchable database.



Conclusion and Future Works

We have successfully completed the data collection and extraction phase, establishing a functional workflow for gathering and processing relevant medical reports. The system is now capable of storing and organizing the extracted information efficiently. With this solid foundation in place, we are moving forward to the model training phase. Our ultimate goal is to develop a user-friendly tool that helps scientists stay continuously informed about the latest developments in their field, especially within the fast-paced, ever-evolving world of medical research.

References

PubMed Key Statistics. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov
Lewis, Pet et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474. https://arxiv.org/abs/2005.11401
Image generation: ChatGPT 4o https://chat.openai.com/
Image editing: Krita: https://krita.org/

Scan Me

Contacts
Al Akl Jihad
+961 70533519 – jihad.alakl@ua.edu.lb

Besançon, Belfort, Montbéliard, Sévenans
www.femto-st.fr